(12) UK Patent Application (19) GB (11) 2 359 398 (13) A

(43) Date of A Publication 22.08.2001

(21) Application No 0004069.1

(22) Date of Filing 21.02.2000

(51) INT CL7
G06F 3/023 , H03M 11/00

(52) UK CL (Edition S)
G4H HKJ H13D H14A H14B H14D
U1S S2127 S2215

(56) Documents Cited
None

(58) Field of Search
UK CL (Edition R) G4H HDQ HDW HKJ
INT CL7 G06F
ONLINE:WPI,EPODOC,JAPIO

(71) Applicant(s)
Kenwood Corporation
(Incorporated in Japan)
14-6,Dogenzaka 1-Chome, Shibuya-ku, Tokyo, Japan

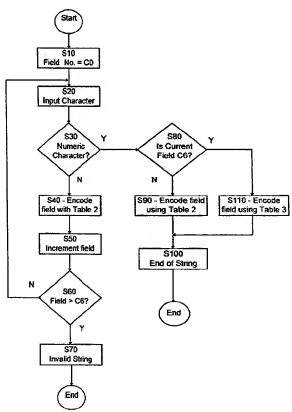
(72) Inventor(s)
Shide Wang

(74) Agent and/or Address for Service
R.G.C.Jenkins & Co
26 Caxton Street, LONDON, SW1H ORJ,
United Kingdom

(54) Abstract Title
Encoding Chinese characters

(57) In a method of encoding a Pinyin string, each of the letters of the Pinyin string is represented by five bits. The final tone digit, which can be in the range of 1 to 5, is also represented by five bits, unless the Pinyin string contains six letters. In Pinyin strings containing the maximum of six letters, the tone digit '5' is never used, so that in this case the final tone digit is represented by only two bits. Therefore, any valid Pinyin string can be represented by a 32-bit word, which can be efficiently stored and compared with other Pinyin strings. The method is particularly suitable for compact, low power text storage and/or messaging devices.

Fig. 3



GB 2 359 398 A

Fig. 1

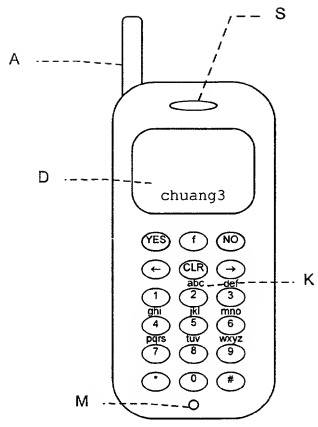

S
A
D
chuang3
YES
f
NO
CLR
abc
def
K
1
2
3
ghi
jkl
mno
4
5
6
pqrs
tuv
wxyz
7
8
9
*
0
#
M

Fig. 2

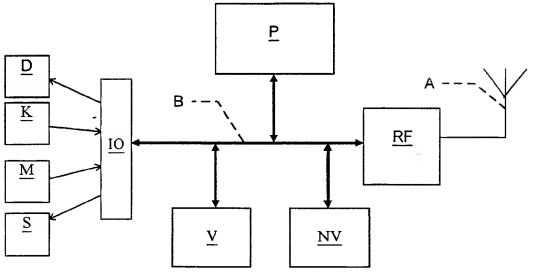
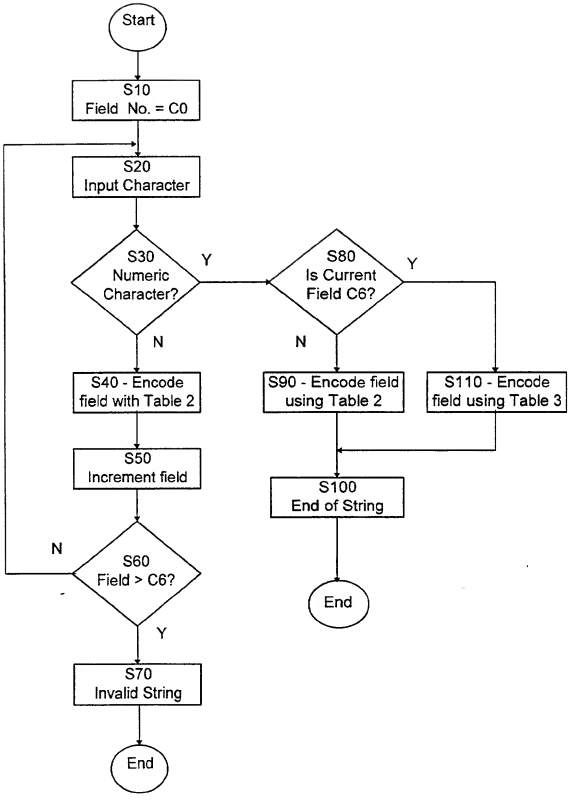

P
D
K
M
S
IO
B
A
RF
V
NV

Fig. 3


Start
S10
Field No. = C0
S20
Input Character
S30
Numeric
Character?
Y
N
S80
Is Current
Field C6?
Y
N
S40 - Encode
field with Table 2
S90 - Encode field
using Table 2
S110 - Encode
field using Table 3
S50
Increment field
S100
End of String
S60
Field > C6?
N
Y
End
S70
Invalid String
End

ENCODING METHOD

TECHNICAL FIELD OF THE INVENTION

The present invention relates to a method of encoding characters, particularly Pinyin representations of Chinese characters, and particularly in embedded systems such as mobile telephone handsets.

BACKGROUND

Pinyin is a Romanized phonetic system used to represent Chinese character pronunciations. Normally, Pinyin strings are input to electronic devices letter by letter and ASCII strings are used by the devices for internal processing. The maximum length of a Pinyin string necessary to represent one Chinese character is 7 bytes (6 letters plus a digit), so that 8 bytes of memory space is required to represent each Pinyin string internally in the electronic device. The final tone digit allows distinction between different pronunciations of the same Chinese character.

There are nearly 7000 Chinese characters per Chinese language and some characters may have up to 5 different pronunciations. Therefore, Pinyin databases can be very large and string comparison very slow. These drawbacks are of little consequence in software for general purpose computers, but can be critical in embedded systems, such as mobile handsets, where processor speed and storage are limited by the power and size constraints of the system.

DISCLOSURE OF THE INVENTION

According to the present invention, there is provided a method of encoding a Pinyin string, in which the string is compressed into a single 32-bit word. This allows compact and rapid storage and handling of the Pinyin strings, particularly in systems with a 32-bit architecture.

Preferably, each of the letters of the Pinyin string is represented by five bits. Preferably, the final digit is also represented by five bits, unless the Pinyin string contains six letters, in which case the final tone digit is represented by two bits. At first sight, two bits appear insufficient to store a digit in the range one to five. However, in Pinyin strings containing the maximum six letters, the tone digit '5' is never used. Therefore, any valid Pinyin string can be represented by a 32-bit word.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a front view of a mobile telephone handset;

Figure 2 is a diagram of the internal electronic components of the handset; and

Figure 3 is a flowchart of an encoding algorithm in an embodiment of the present invention.

MODES FOR CARRYING OUT THE INVENTION

Figure 1 shows a mobile telephone handset H having a keypad K comprising numeric keys 0 to 9, star (*) and hash (#) keys, and function keys such as 'YES', 'NO', back/up (←), forward/down (→), clear (CLR) and other function (f). A display D is able to display Arabic numerals, Roman letters and Chinese characters, and may be an LCD with sufficient resolution to display at least one line of numerals, letters and characters. A microphone M and speaker S are also present, to allow voice calls.

Figure 2 is a schematic diagram of the electronic components of the handset H. These components need not be discrete, and may be integrated. For example, the components may be integrated onto a microcontroller chip and an RF stage chip. A processor P is connected via a bus B to a volatile memory (V), a non-volatile memory (NV), an I/O interface (I/O) and an RF modem (RF). The I/O interface decodes input from the keypad K and microphone M

and drives the display D and speaker S. The RF modem is connected to an antenna A so as to receive and transmit RF signals. The components are powered by a battery (not shown) or a mains electricity connection (not shown) via a transformer.

The non-volatile memory (NV) stores software which is executed by the processor P in order to carry out the functions of the handset. Optionally, the non-volatile memory may be reprogrammable to upgrade the software. The algorithm described below as an embodiment of the present invention may be installed as an upgrade to the software of an existing mobile telephone. The upgrade may be received as a wireless message. via the RF modem (RF).

The handset H implements protocols which allow text messages to be sent and received. For example, the handset may be GSM-compatible and support the GSM SMS (short message service) protocols. As is generally known, alphanumeric characters are entered on the keypad K by a predetermined sequence of key presses. For example, as shown in Figure 1, more than one Roman letter is assigned to each numeric key and the appropriate letter is selected by multiple short presses of the same key until the desired character is displayed. Characters not displayed on the numeric keys, such as punctuation, may be selected by further rapid key presses. Alternative character selection methods are also known, such as predictive input in which the user need normally only press each key once for each letter and the software running on the handset guesses which combination of letters is intended by comparing possible combinations of letters with valid words stored in memory, for the language selected by the user.

In an embodiment of the present invention, the user can enter Pinyin strings for transmission as text messages via the mobile radio network with which the handset H operates, by selecting a Pinyin entry mode. The user spells out the Pinyin string using the keys of the keypad K, using any of the

known techniques for entering Roman letters and Arabic numerals on a numeric keypad. Entry of Pinyin letters is not case sensitive. While the individual letters of the Pinyin string are being entered, they are displayed on the displayed and stored in the volatile memory (V). For example, the user may enter the Pinyin string 'chuang3'.

Software running on the handset H identifies when a complete Pinyin string has been entered. For example, when a numeral is entered, it can be assumed that this is the last character of a Pinyin string. Alternatively, there may be stored in the non-volatile memory a database of all valid Pinyin strings and the software may display the equivalent Chinese character when the characters of the Pinyin string entered by the user are sufficient to identify one character uniquely, or display all possible Chinese characters when all the possible characters can be displayed. The user may be required to press another key to confirm that the string entered is correct, or to select the intended Chinese character if there is more than one possibility.

The software encodes each completed Pinyin string as a 32-bit word and this encoded form is preferably used for storage of Chinese text prior to transmission, and for the storage of any databases of Pinyin strings on the handset, for example in the non-volatile memory. Such databases may be used for predictive input or for validation of entered Pinyin strings, as described above.

The format of the 32-bit word is shown below in Table 1:

Table 1

Bit nos.	Field
31-27	C0
26-22	C1
21-17	C2
16-12	C3
11-7	C4
6-2	C5
1-0	C6

The following rules are applied for compressing each Pinyin string:

1) Left alignment: the first Pinyin letter is in Field C0, the second in C1 and so on.

2) Any unused fields are set to all zero bits.

3) If the length of the Pinyin string is less than or equal to six characters (five letters and one digit), then the letters are encoded in each of the fields C0 to C5 as in Table 2 below:

Table 2

Pinyin Character	Binary code
1	00001
2	00010
3	00011
4	00100
5	00101
a	00110
...	...
z	11111

4) If the length of the Pinyin string is seven characters (six letters and one digit), then the fields C0 to C5 are encoded as in Rule 3 and Table 2 above. However, field C6 is encoded as shown below in Table 3:

Table 3

Pinyin Tone Digit	Binary code
1	00
2	01
3	10
4	11

The tone digit 5 does not need to be encoded, because there are no six-letter Pinyin strings having tone digit 5. Therefore, no information is lost in the compression algorithm given above.

As an example, the Pinyin string 'chuang3' is encoded as the binary 32-bit word:

'01000/01101/11010/00110/10011/01100/10'

where the slash character demarcates the fields C0 to C6, but does not represent any additional bits.

The encoding algorithm for each Pinyin string can be represented as a flowchart as shown in Figure 3. At step S10, the field number is set to C0. At step S20, a character is entered. As part of this step, the character may be checked to ensure it is an acceptable Pinyin string character and the step may continue until a valid character is entered. At step S30, it is determined whether the entered character is a numeral. If not, at step S40 the code value of the entered character, according to Table 2 above, is entered in the current field. At step S50, the current field number is incremented. At step S60, it is determined whether the current field number exceeds C6. If so, the maximum Pinyin string length has already been reached without a numeral being

entered, so it is indicated at step S70 that the string is invalid. If not, the flow returns to step S20.

If at step S30 it is determined that the entered character is a numeral, it is then determined at step S80 whether the current field number is C6. If not, at step S90 the code value of the entered numeral, according to Table 2 above, is entered in the current field, and the end of the string is indicated at step S100. If the current field is C6, the code value of the entered numeral, according to Table 3 above, is entered in that field at step S110, and the end of the string is indicated at step S100.

Alternatively, the encoding steps may take place only after a complete Pinyin string has been input. Moreover, the Pinyin string may be checked against a database of valid Pinyin strings and the user prompted to edit the string if it is not valid, or the closest matches may be displayed to the user for selection. Preferably, the entered Pinyin string is encoded as a 32-bit word and compared with a database of valid Pinyin strings also encoded as 32-bit words. The processor P is typically able to handle 32-bit words as integers which can be fetched from memory in a single operation, and may have an instruction set which includes a single instruction to compare 32-bit integers. Hence, the comparison of the entered Pinyin string with a database of Pinyin strings may be performed much more quickly than by performing a string comparison between uncompressed strings. The encoded Pinyin strings may be stored much more compactly than the equivalent ASCII strings. Hence the compression algorithm is particularly suitable for implementing search and storage of Pinyin strings on a compact, low-power device.

The above description relates to a mobile telephone but it will readily be understood that the compression algorithm is equally suitable for text-only transceivers or PDA's (personal digital assistants).

CLAIMS

1. A method of encoding a Pinyin string comprising a plurality of Roman letters and one numeral to generate an encoded Pinyin string, including:

encoding each said Roman letter using a constant number of bits, and, if there are six of said Roman letters in the Pinyin string, encoding said numeral in two bits.

2. A method as claimed in claim 1, wherein, if there are less than six said Roman letters in the Pinyin string, the numeral is encoded using said constant number of bits.

3. A method as claimed in claim 1 or claim 2, wherein said constant number is five.

4. A method as claimed in any preceding claim, wherein the encoded Pinyin string has a length of 32 bits.

5. A method of searching a database of Pinyin strings, each encoded by the method of any preceding claim, including:

encoding a search string including one or more Pinyin strings, by means of a method according to any preceding claim;

comparing said encoded search string with some or all of said database of Pinyin strings; and

indicating, on the basis of said comparison, whether the encoded search string matches any of said database of Pinyin strings.

6. A method of storing a plurality of Pinyin strings, comprising:

encoding each of said plurality of Pinyin strings by means of a method as claimed in any one of claims 1 to 4, and

storing the encoded Pinyin strings in a memory.

7. Apparatus arranged to perform a method as claimed in any preceding claim.

8. A portable electronic device including apparatus as claimed in claim 7.

9. Software arranged to perform a method as claimed in any one of claims 1 to 6.

10. A signal containing one or more encoded Pinyin strings encoded by a method as claimed in any one of claims 1 to 6.

11. A method substantially as herein described with reference to Figure 3 of the accompanying drawings.

Application No:	GB 0004069.1	**Examiner:**	Mike Davis
Claims searched:	1-11	**Date of search:**	4 May 2000

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.R): G4H (HKJ, HDW, HDQ)

Int Cl (Ed.7): G06F

Other: Online: WPI, EPODOC, JAPIO

Documents considered to be relevant:

Category	Identity of document and relevant passage	Relevant to claims
	None	

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.